DELAY-SENSITIVE COMMUNICATION OVER WIRELESS MULTIHOP CHANNELS

A Thesis

by

OMAR AHMED ALI

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

May 2008

Major Subject: Electrical Engineering

DELAY-SENSITIVE COMMUNICATION OVER WIRELESS MULTIHOP CHANNELS

A Thesis

by

OMAR AHMED ALI

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Approved by:

Chair of Committee,	Jean-François Chamberland
Committee Members,	Krishna R. Narayanan
	Jim Xiuquan Ji
	Natarajan Gautam
Head of Department,	Costas N. Georghiades

May 2008

Major Subject: Electrical Engineering

ABSTRACT

Delay-Sensitive Communication over Wireless Multihop Channels. (May 2008)

Omar Ahmed Ali, B.Tech, Indian Institute of Technology Madras

Chair of Advisory Committee: Dr. Jean-François Chamberland

Wireless systems of today face the dual challenge of both supporting large traffic flows and providing reliable quality of service to different delay-sensitive applications. For such applications, it is essential to derive meaningful performance measures such as queue-length distribution and packet loss probability, while providing service guarantees. The concepts of effective bandwidth and effective capacity offer a powerful cross-layer approach that provides suitable performance metrics for the bandwidth and capacity of wireless channels supporting delay-sensitive traffic. Many wireless systems rely on multihop forwarding to reach destinations outside the direct range of the source. This work extends part of the methodology available for the design of wireless systems to the multihop paradigm. It describes the analysis of a communication system with two hops using this cross-layer approach. A framework is developed to study the interplay between the allocation of physical resources across the wireless hops and overall service quality as defined by a queueing criterion based on large deviations. Decoupling techniques introduce simple ways of analyzing the queues independently. Numerical analysis helps identify fundamental performance limits for Rayleigh block fading wireless channel models with independent and identically distributed blocks. Simulation studies present comparable results akin to that obtained using the analytical framework. These results suggest that it is imperative to account for queueing aspects while analyzing delay-sensitive wireless communication systems.

To My Parents and Sisters

ACKNOWLEDGMENTS

First and foremost, I would like to thank God Almighty for giving me an opportunity to come to the United States and study in one of the most prestigious universities. The knowledge and experience I have gained here will provide a strong foundation for my future career, and I am confident that it will help me conduct His work.

Next, I wish to express my sincere and heartfelt gratitude to my advisor Dr. Jean-François Chamberland for his valuable guidance, time and constant encouragement. I thank him for the confidence he had in me in granting me this project, and for the countless other times I knocked on his door asking for assistance.

I am grateful to my colleagues Parimal, Lingjia, Abdallah, Arvind and Nirmal for the priceless discussions and brain-storming sessions, which have helped me in completing this thesis successfully.

I appreciate all the esteemed professors who have taught me and inspired me throughout my graduate studies. I thank all my friends at Texas A&M University who have made my stay here the most memorable time of my life. Finally, I would like to thank my family and relatives who have always supported me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

FIGURE Page

CHAPTER I

INTRODUCTION

In the recent past, there has been an increasing demand for wireless network access throughout the world. This demand has paved the way for the vast ongoing research in wireless technologies. Wireless networks face the dual challenge of supporting large traffic volumes and providing reliable service for delay-sensitive applications such as Voice over Internet Protocol (VoIP), wireless security systems, video conferencing, electronic commerce, sensor networks and gaming. Most of the research on wireless systems available in the literature today focuses on maximizing physical layer attributes. For instance, several papers are concerned with computing the Shannon capacity [1] and spectral efficiency [2, 3] associated with specific wireless schemes. However, real-time applications typically have stringent service requirements for which a classical capacity-based analysis does not offer a complete assessment of service quality. This is especially true for the communication infrastructures associated with wireless networks, as they are subject to time-varying service. A study of queueing dynamics is essential to relate the effects of decisions at the physical layer to the service requirements of delay-sensitive networks [4]. Performance measures such as queue length, packet loss probability, and delay influence the perceived quality of a communication link. Requirements on these attributes may force a wireless system to operate well below its theoretical Shannon limit. These measures should be taken into account while designing wireless systems. Hence, to analyze such systems accurately, we favor a framework that utilizes cross-layer design techniques by exchanging information between the physical layer and the data-link layer.

The journal model is *IEEE Transactions on Automatic Control.*

Quality of service (QoS) has been studied extensively in the context of wired networks [5, 6, 7]. Due to the time-varying nature of real-time sources and wireless channels, it is quite difficult to provide deterministic delay guarantees to mobile users. Several models for communication over wireless channels have been proposed to study the tradeoffs between average transmit power and average queueing delay [8]. These models employ the concepts of outage probability [9] and delay-limited capacity [10] to characterize these tradeoffs. The problem of minimizing queueing delay for a time-varying channel with a single queue, subject to constraints on average and peak power is studied in [11]. Recently, studies have been conducted to obtain throughput optimal control policies for cooperative relay networks that take queue dynamics into account [12]. Hence the use of cross-layer techniques in analyzing wireless networks is gaining momentum because of their significance to real-time applications. Delay violation probability is a better statistical performance measure than average queueing delay in terms of service requirements for many real-time applications. For example, consider a system with VoIP application having a maximum delay tolerance of 300 ms. There could be a case in which the average delay of all the voice-packets in the system is well below 300 ms while most of the packets reach the destination after the allowed delay tolerance. The average delay in this case would be a bad metric for measuring performance of the system. The delay violation probability is defined as the probability of a delay bound being transgressed. Early research on statistical performance guarantees for time-varying traffic led to the unifying concept of effective bandwidth [13]. Given a specific arrival process, the effective bandwidth characterizes the minimum data-rate (bandwidth) required for the communication system to meet a certain QoS requirement. The concept of effective bandwidth was later extended to effective capacity for wireless systems [14, 15, 16]. Assuming a constant flow of incoming data, the effective capacity characterizes the maximum arrival rate that a

wireless system can support subject to a QoS requirement. Both these concepts are useful tools to identify system limitations as a function of statistical queueing violation probabilities. In general, the delay violation probability of a queueing system can be upper bounded in terms of the probability of buffer overflow [17]. A statistical QoS metric that underlies the concept of effective bandwidth and which has been adopted widely in literature is the asymptotic decay rate of buffer occupancy [13],

$$\theta = -\lim_{x\to\infty} \frac{\log \Pr(L > x)}{x} \tag{1.1}$$

where L is the steady-state queue-length distribution of the buffer present at the transmitter. Parameter θ captures the perceived quality of a communication link, and partly reflects user satisfaction. A larger θ implies that the queue length exceeds the overflow threshold with a very low probability, hence representing a more reliable connection or a tighter QoS constraint. There has been recent work done to characterize the interplay between the physical layer infrastructure and the queueing behavior of a single-hop wireless system using this evaluation framework [18].

Still, many wireless systems rely on multihop forwarding to reach destinations outside the direct range of the source. Thus we need to characterize the overall effective bandwidth and capacity of multihop systems to analyze their performance under service constraints. In applications such as wireless sensor networks, there are strict constraints on the amount of power that a sensor can use to transmit its information and hence other sensors in the vicinity are used to relay or route this data to the cluster heads. In addition, the end-to-end delay must be maintained as per user service requirements. A detailed performance analysis based on QoS requirements is yet to be completed for physical channels corresponding to networks consisting of multihop connections. This has been the motivation behind the work described in this thesis.

Specifically, we use the cross-layer approach briefly described above to investigate the queueing behavior of wireless communication systems with multihop links. We employ the effective capacity to assess the overall performance of the system and devise design guidelines for wireless communication systems. The analysis framework developed herein is flexible enough to accommodate small hop-counts. We note that classical network calculus [19, 20] cannot be applied to this scenario due to the time-varying nature of wireless channels. Rather, the exposed research introduces decoupling techniques that provide upper and lower bounds on overall performance.

A. Problem Statement

Consider a simple wireless communication system where data gets transmitted via two hops before it reaches the destination, as shown in Figure 1. The overall system is subject to a mean power constraint and a finite spectral bandwidth allocation. Suppose that a large buffer is available at every transmitter where outgoing packets are stored before being sent to their destination. We will consider a case in which each queue in the system must satisfy a QoS constraint θ as defined in (1.1). Finally, we also assume that channel state information is not available at the transmitters, although the channel statistics are. Our goal is to study the relationship between the allocation of physical resources across the different links and the performance of the system in terms of its QoS constraint on queueing behavior.

B. Thesis Structure

Chapter II illustrates the challenges faced in the design and implementation of wireless networks with delay-sensitive applications. It also presents the mathematical tools available in literature that will be used to develop an analysis framework for

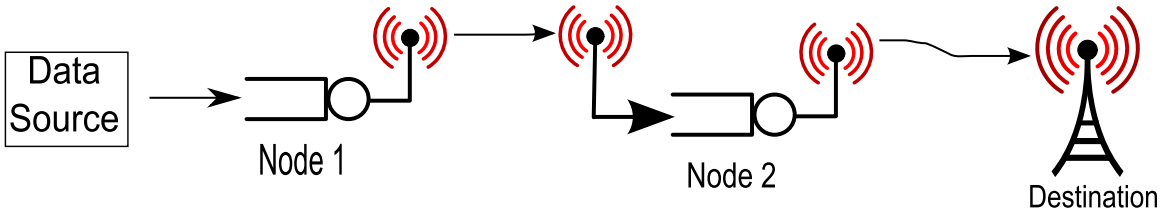

Fig. 1. Abstract model of a communication system with two hops.

the problem at hand. Decoupling techniques used to analyze multiple-queue systems are then introduced. Chapter III describes the two-hop communication system model we adopt. The Rayleigh block fading model used for representing the wireless channels and the discrete-time queueing models are explained elaborately. Chapter IV analyzes the performance of the queueing system. An effective capacity based study is conducted to relate effects of the QoS requirement on the optimal allocation of physical resources. Chapter V presents the results obtained from the simulations of the actual two-hop system, and elucidates how they relate to the analysis framework developed. Conclusions and the scope of future work are presented finally.

CHAPTER II

WIRELESS SENSOR NETWORK AND MULTIHOP CHANNELS

A wireless sensor network is a data-acquisition system consisting of spatially distributed autonomous devices using sensors to cooperatively monitor physical or environmental conditions, such as temperature, sound, vibration, pressure, motion or pollutants, at different locations [21]. The development of wireless sensor networks was originally motivated by military applications such as battlefield surveillance. However, wireless sensor networks are now used in many civilian application areas including environment and habitat monitoring, pollution control, health-care applications, home automation, and traffic control. In a typical wireless sensor network architecture, packets reach cluster heads in a small number of hops as shown in Figure 2.

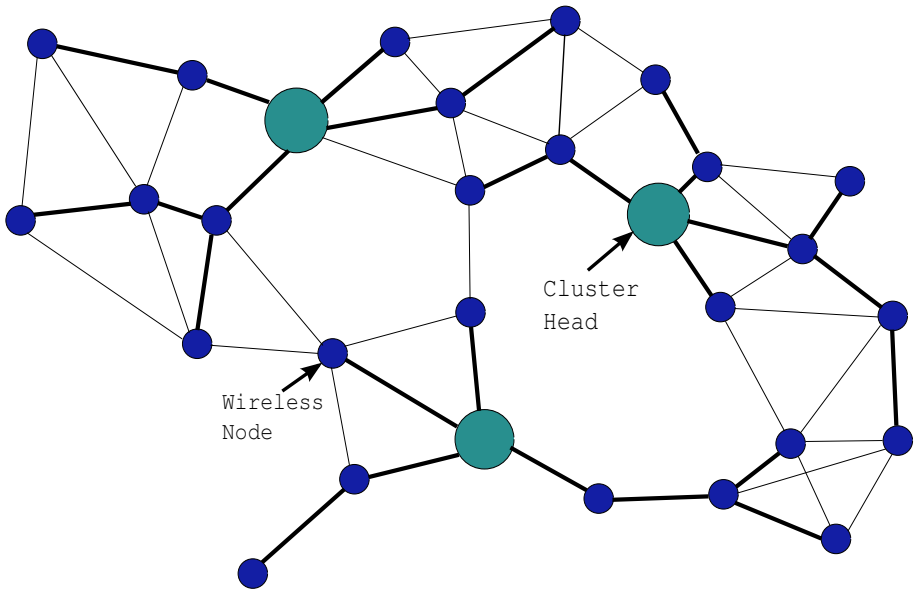

Fig. 2. A generic wireless sensor network architecture.

In addition to having one or more sensors, each node in the system is typically

equipped with shoe-box-sized a radio transceiver or another wireless communications device, a small micro-controller, and a power source, usually a battery. The volume of a sensor node can vary from shoe-box-sized nodes down to devices the size of a grain of dust. The cost of individual sensor nodes is similarly variable, ranging from hundreds of dollars to a few cents, depending on the purpose of the sensor network and the complexity required of each node. Size and cost constraints on sensor nodes result in corresponding constraints on resources such as power, memory, computational speed and bandwidth.

Key to the success of wireless sensor technology is a communication strategy that is tailored to distributed signal processing. The protocols used should enable efficient, adaptive and robust data transfers within the network. Wireless sensor networks cannot be viewed as extensions of ad-hoc wireless networks with stringent power and energy constraints because the focus in conventional data networks is on connectivity and improving throughput [22], independent of the applications. Whereas in sensor networks, the optimization metrics are formulated primarily by the underlying sensing objectives such as probability of detection error, detection delay or mean-square error in estimation [23, 24, 25]. Hence the design process for a sensor network should be closely knit to its application objective and associated performance metric.

The performance of a wireless sensor network is governed by its ability to gather, carry and use relevant information in a timely fashion. Since wireless nodes are typically subject to strict power constraints, it is essential to study and understand the interplay between resource allocation and overall performance in such systems [26, 27, 28]. A major challenge in the design and implementation of wireless sensor network revolves around distributed inference problems and the transmission of delay-sensitive traffic over the network. This challenge can be met by adopting a cross-disciplinary approach incorporating ideas spanning communication at the phys-

ical layer, statistical signal processing, queueing theory and networks. Research on wireless sensor networks can be organized around three interdisciplinary areas. The first area involves the detection of events at the sensor nodes and estimation of the state of nature at the fusion center [29, 30, 31]. Proper placement of nodes, sensor density and efficient quantization of the data locally play important roles in determining how this decentralized inference problems can be handled. Providing reliable communication schemes for delay-sensitive data transfers is another crucial interdisciplinary area for the design of sensor networks. This aspect is covered in detail in Section B of the present chapter. The third field of research is the development of distributed protocols for information-aware sensor networks. An information-aware sensor network is a system that gives priority to packets that contain pertinent information. Knowledge of the quality of data contained in the packets can be used to devise effective communication strategies, scheduling schemes and routing protocols. Similarly sensor nodes can conserve energy by having a priori knowledge about the process they are monitoring together with their current and past observations.

Techniques from large deviations [32, 33] can be employed for the design of resource-constrained wireless sensor networks. In the past, large-deviation theory has been used extensively in the context of inference problems [34, 35, 36] and network analysis [7, 37, 38]. Techniques extracted from the literature on large deviations are particularly meaningful in the context of sensor networks because they capture the dominating behavior of large systems. These techniques can be leveraged to derive meaningful performance metrics for the design of delay-sensitive wireless sensor networks, analyze prominent interference patterns, compute optimal node density and placement, and provide statistical delay guarantees for data transmissions.

A. Large Deviation Principle

The large deviation principle (LDP) plays a fundamental role in both information theory and queueing theory and, in many cases, it provides tight characterizations of system performance. The LDP is closely related to the distinguished concept of error exponent for random codes. For discrete memoryless channels, error exponent is tight for rates sufficiently close to capacity [39]. The LDP captures the asymptotic behavior, as $\epsilon \to 0$, of a collection of probability measures $\{\mu_\epsilon\}$ in terms of a rate function. The set of measures $\{\mu_\epsilon\}$ satisfies the LDP with rate function $I(\cdot)$ if, for every admissible set S,

$$-\inf_{s\in\text{interior}(S)} I(s) \le \liminf_{\epsilon\to 0} \epsilon \log \mu_\epsilon(S) \le \limsup_{\epsilon\to 0} \epsilon \log \mu_\epsilon(S) \le -\inf_{s\in\text{closure}(S)} I(s).$$

For instance, let M_n be the empirical average of n zero-mean independent and identically distributed (i.i.d.) random variables $X_1, X_2, \ldots, X_n$, each with finite second moment. The weak law of large numbers asserts that M_n converges to zero in probability. The LDP states that the tails of the distribution of M_n decay exponentially fast, $\Pr(M_n > t) \asymp \exp(-nI(t))$ and $\Pr(M_n < -t) \asymp \exp(-nI(-t))$ for $t > 0$. In this particular case, the rate function $I(\cdot)$ is the Fenchel-Legendre transform of the cumulant generating function of X. The Fenchel-Legendre transform of a log moment generating function $\Lambda(\theta)$ is defined by

$$\Lambda^*(t) = \sup_{\theta\in\mathbb{R}} \{\theta t - \Lambda(\theta)\}.$$

The large deviations principle is closely related to the Chernoff bound. Suppose $\phi(x) = e^{\theta x}$ with $\theta > 0$, and let $S = \{s : s \ge t\}$ where $t > 0$. Then, using Markov's inequality, we get

$$e^{\theta t} \Pr(X \ge t) \le E\left[e^{\theta X}\right],$$

where $E\left[e^{\theta X}\right]$ is the moment generating function of random variable X. Taking natural logarithms of both sides, we obtain

$$\log \Pr(X \geq t) \leq -\theta t + \log E\left[e^{\theta X}\right].$$

The probability that the empirical mean of the first n variables exceeds $t > 0$ is bounded by

$$\Pr(M_n \geq t) = \Pr\left(\sum_{i=1}^{n} X_i \geq tn\right) \leq e^{-\theta tn} E\left[e^{\theta \sum_{i=1}^{n} X_i}\right] = \left(e^{-\theta t} E\left[e^{\theta X}\right]\right)^n.$$

It follows that, for any $\theta \geq 0$, we have

$$-\frac{1}{n} \log \Pr(M_n \geq t) \geq \theta t - \log E\left[e^{\theta X}\right] = \theta t - \Lambda(\theta).$$

Maximizing the right hand side over θ, we get

$$-\frac{1}{n} \log \Pr(M_n \geq t) \geq \sup_{\theta \geq 0}\{\theta t - \Lambda(\theta)\} = \Lambda^*(t).$$

A derivation showing that the Chernoff bound is asymptotically tight can be found in [32]. The purpose of this brief discussion is to help the reader gain intuition about good rate functions and LDPs. For an LDP to apply, sequences of random variables need not be i.i.d. The Gärtner-Ellis theorem provides sufficient conditions for a rate function to exist [32].

Sample path large deviations is a closely related concept on which many network performance metrics are derived. In queueing systems, explicit expressions for error probability, delay distribution, and queue length are difficult to get. The theory of sample-path large deviations is therefore frequently employed to characterize the dominating behaviors governing queue-length distributions. Consider a single-server queue whose distribution obeys an LDP. Its rate function may be dominated by the arrival process, the service process, or by joint deviations in both. When the service

process of the queue is determined by a wireless channel, the LDP of the queue-length depends on the statistics of the wireless channel, which are closely related to the error exponent of the channel. Large deviations form a basis for the effective capacity and effective bandwidth, two performance metrics introduced in Chapter I.

A systematic application of large deviations to various fields and an introduction to the theory can be found in [40]. One may find in the literature results more precise than the large deviation principle (LDP). Although it can be argued that the LDP provides only some rough information on asymptotic probabilities, its scope and ease of application have made it a popular tool [32]. The theory of large deviations has a beautiful and powerful formulation due to Varadhan, along with Chernoff's theorem, making it very general [41]. More information about sample path large deviations can be found in [32].

B. Delay-Sensitive Communication

To capture the overall performance of a delay-sensitive wireless network, an analytic study of the queueing behavior is essential in addition to the physical layer concept of Shannon capacity. Based on the concepts from large-deviation theory discussed earlier, an evaluation methodology suitable to characterize system limitations under queueing constraints can be constructed. Such a framework would lead to achievability results akin to Shannon capacity, albeit in a QoS framework. This section motivates the problem addressed in this thesis by briefly describing different problems in wireless communication systems that have been studied using the concepts of effective bandwidth and effective capacity.

A performance analysis of a single-hop Gilbert-Elliot wireless system was conducted as a function of physical resources using the probability of buffer overflow as

a QoS constraint in [18]. In their work, the authors show that the effective capacity decays sharply as a function of the QoS constraint $\theta > 0$ defined in (1.1). In addition, it is shown that the optimal code selection for a wireless system depends on its QoS requirement. A more stringent constraint on θ lowers the optimal coderate. Correlation of the underlying physical channel is also found to have a major impact on performance. The effective capacity of a slowly varying channel can be very small. It was shown that, even with unlimited amount of physical resources, the maximum arrival rate supported under QoS constraint θ is bounded.

This cross-layer framework was also employed to analyze the benefits of user-cooperation in delay-sensitive wireless systems; and the corresponding achievable rate-region for a two-user scenario is characterized in [42]. Numerical results suggest that cooperation yields a large gain over traditional paradigms. User-cooperation can therefore provide wireless users with the flexibility to better share system resources. Again, the overall performance seems to depend heavily on the time-correlation of the physical channel. This emphasizes the fact that effective capacity is much more sensitive to higher-order statistics than ergodic capacity or outage capacity.

System performance for a class of multiple-antenna wireless systems subject to Rayleigh flat fading is studied [17]. The effective capacities of various vector Gaussian channels are characterized, and overall performance is evaluated in the low signal-to-noise ratio regime. Moreover, the dominating behaviors of MIMO systems are analyzed in the large antenna-array regime. Numerical results confirm that the potential gains of multiple-antenna configurations over single-antenna systems are substantial. When the number of transmit and/or receive antennas becomes large, the effective capacity of the system is bounded away from zero even under very stringent service constraints. This phenomena, which results from channel-hardening, suggests that a multiple-antenna configuration is greatly beneficial to delay-sensitive traffic.

Recent work has been done to compare the performance of a butterfly network with and without network coding in the context of delay sensitive applications [43]. The butterfly network is a system in which the information from multiple sources can be transmitted to multiple destinations, with each destination being able to recover messages from all the sources. Information can be transmitted through intermediate nodes which either route packets or perform algebraic manipulations on the locally available packets. The achievable rate-region was computed for a butterfly network with two sources and two destinations, when operating under strict QoS requirements and in the context of both wired and wireless communication. For a wireless network, combining packets at intermediate nodes doesn't necessarily offer performance gains. Rather, in some cases, it can be harmful. Hence in these instances, it is better to just route the packets.

It is quite evident that using a cross-layer approach for the analysis of delay-sensitive wireless communication systems may provide better insight into the actual working and performance of such systems. A number of systems rely on multihop transmissions to send data to remote destinations. As such, it is imperative that multihop systems be analyzed using a similar framework. A QoS analysis of multihop systems can provide us with an understanding of how nodes should be distributed in a wireless sensor network in order to maximize performance.

In reality, it is difficult to analyze a complete multihop system because departure processes of queues are hard to characterize. Thus, we approximate their behaviors using the service processes. Decoupling buffers from one another provides a simple way to analyze the queues independently.

1. Decoupling Techniques

We first state the buffer decoupling argument for a single queue system [18], and then extend it to a multihop scenario. The argument compares a single-queue system to two other systems, each with two queues. This natural progression appears in Figure 3. We begin with a comparison of the first two systems. The arrival process of the second system is assumed to be identical to the arrival process of the first system. Similarly, the service process of the second queue is made equal to the service process of the first system. In addition, the first queue in System 2 is serviced at a constant rate whenever it is non-empty. The packets departing from the first queue are immediately placed in the following queue. The additional constraint present in the latter scenario causes the queue-length in the single-queue system to be always less than or equal to the sum of the queues in the second system.

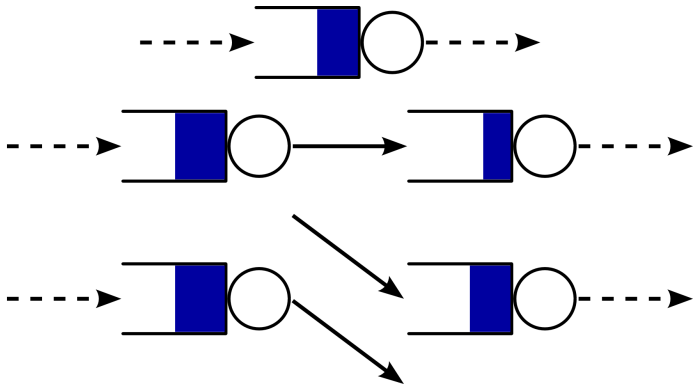

Fig. 3. Illustration of the three systems used in buffer decoupling argument.

Next, we compare the second system with a network composed of two independent queues. The arrival process in the first queue is the same as before, and this queue is served at a constant rate when it is nonempty. Packets arrive in the second queue at the same constant rate they leave the first queue. However, the arrival in

the second buffer is independent of the first buffer and it never goes idle. Packets in the second queue are served at a rate equal to the service rate of the second queue in System 2. Note that the length of the first queue in the third system is always equal to the length of the first queue in the second system. Furthermore, the length of the second queue in the second system is less than or equal to the length of the second queue in the third system. It follows that the large deviation principle governing the queue length in the first system is always less than or equal to the large deviation principle governing the sum of the queues in the third system.

Now, we consider a more elaborate system composed of two queues in tandem. Let the arrival and service processes of the first queue be denoted by $a_1(t)$ and $s_1(t)$, respectively. The departure process of the first queue is fed directly into the second queue, which is served at rate $s_2(t)$. First, we apply the decoupling technique described above to the two queues individually, as shown in Figure 4. The resulting system then features two queues in place of the original ones. The arrival process for the first queue is $a_1(t)$ and this queue is served at a constant rate ν_1. Packets arrive in the second queue at a constant rate ν_1, and they are served at a rate $s_1(t)$. Appropriate choice of ν_1 will eventually determine how close this approximation is compared to the original system. If ν_1 is equal to the effective bandwidth of $a_1(t)$ and effective capacity of $s_1(t)$, then the queue-lengths of the two-queue approximation and that of its original counterpart will possess the same LDP. Departed packets from the second queue are transmitted through the wireless channel before being placed in the third queue, which is served at constant rate ν_2. Packets arrive in the fourth queue at a constant rate ν_2, and they are serviced at rate $s_2(t)$. An argument similar to that of ν_1 can be stated for ν_2 as well. Based on the additional constraints present in this new scenario, it follows that the sum of the queues in the original system is always less than or equal to the sum of the queues in the latter network.

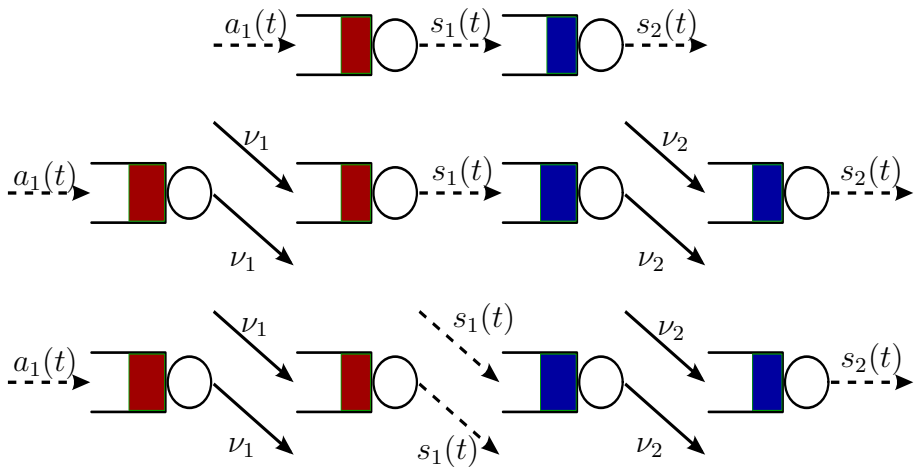

Fig. 4. Decoupling technique applied to a system with two queues in tandem.

Next, we link the second system to a third one where the arrival in the third queue is a stochastic process whose arrival rate is matched to $s_1(t)$, the service process of the second queue. The main difference between the two systems is that the arrival process in the third queue is the departure process of the second queue in one case, whereas it is a process equivalent to the offered service by the second queue in the other one. This approximation again gives an upper bound on the total queue-length of the system. This system can be used to derive performance bounds on the actual system behavior based on appropriate assumptions. Upper bound on the delay can be calculated based on the assumption that the arrival rate is constant and equal to the expected value of the traffic generated by the transmitting queue. But this is a loose bound as in the case when there is no instantaneous traffic being transmitted by the first queue, assuming constant arrival rate at the second queue would be far from accurate. Hence, these bounds become tighter under heavy traffic limits.

This analysis framework can only provide loose bounds in the wired world since in this situation the stochastic nature of the service process is due to priority users rather than the channel variations. However, based on the stochastic nature of wire-

less channels, tighter bounds can be obtained by using appropriate resource allocation policies to smooth out the variance in the channel, possibly at the expense of throughput. The independence structure introduced by the decoupling techniques enables us to analyze the overall two-hop system by studying the two queues independently under proper assumptions.

CHAPTER III

SYSTEM MODEL

The system model consists of wireless links which connect the mobile device and the base station. Each link has three major components: the data stream arriving in the buffer, the length of the buffer, and the service offered to the user. This is shown in Figure 5. The buffer stores the data before being transmitted to ensure that each user receives its intended packets. Yet, the introduction of a buffer adds delay to the transmission procedure, and hence creates a disparity between throughput optimality and delay optimality. We will use the asymptotic decay rate of buffer occupancy introduced in (1.1) as our measure of service quality and delay-sensitiveness.

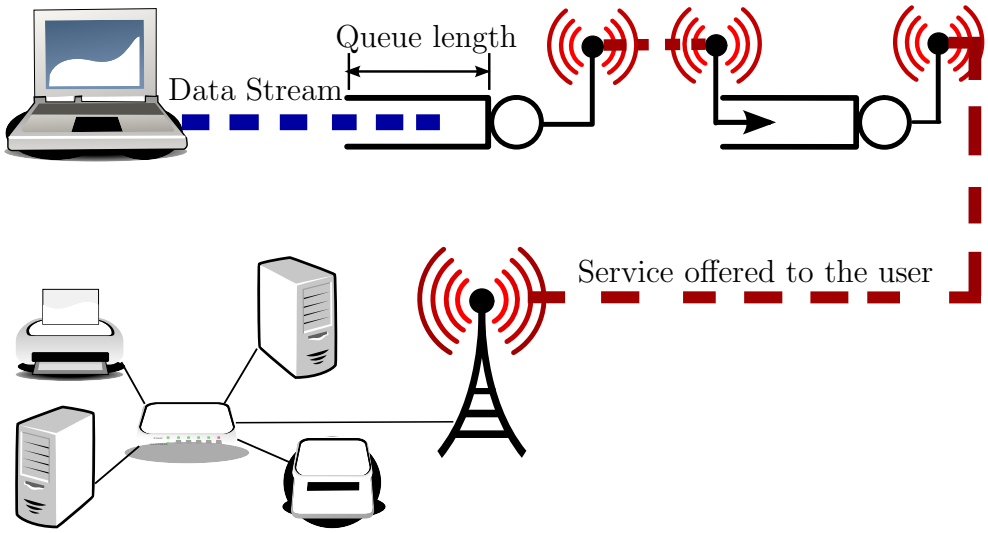

Fig. 5. System model.

To characterize the overall performance of a multihop system in terms of effective capacity, we need to formulate the problem in a queueing theory framework by developing accurate models for the components described above. For wireless applications, the offered service depends on the instantaneous transmission rates supported

by the wireless link. Accurately modeling the behavior of the wireless channels and specifying its evolution in terms of physical layer parameters is essential. This is accomplished in the sections that follow.

A. Multipath Fading

In wireless environments, RF signals from the transmitters may be reflected from objects such as hills, buildings, or vehicles before reaching their destinations. The superposition of these multiple transmission paths at the receiver gives rise to a phenomenon known as multipath fading. Figure 6 shows some of the possible ways in which multipath fading can occur. The relative phase of multiple reflected signals can

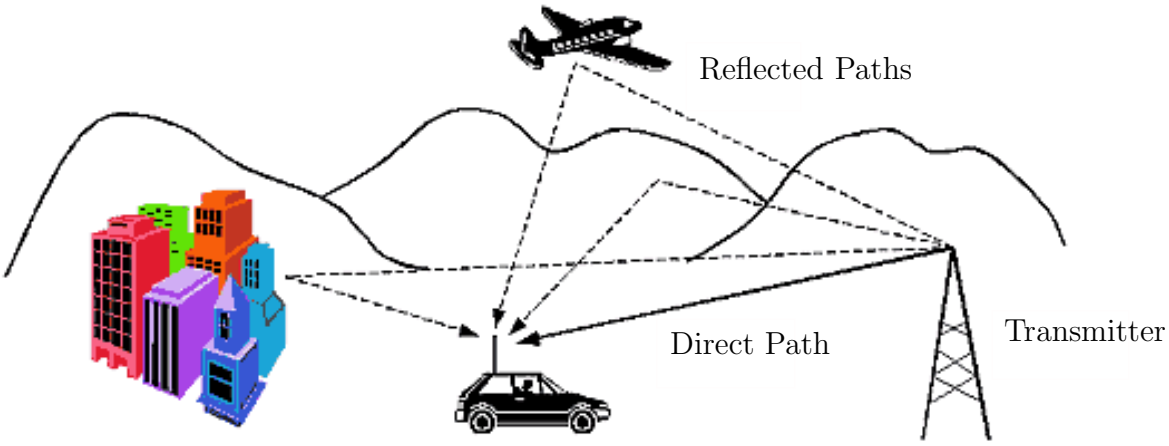

Fig. 6. Multipath signals.

cause constructive or destructive interference at the receiver, depending on the nature of the wireless environment and the mobility of the terminals. Detailed information about channel modeling for wireless communication can be obtained from [44, 45, 46]. For the sake of our discussion, we only briefly describe the pertinent models for fading channels that are often used in the analysis of wireless systems [47].

A simple, yet powerful abstraction for a wireless environment consists of representing the channel as a linear time-varying system. Let $x(t)$ be the signal transmitted over the channel and let $X(f)$ denote its frequency content. Then, the corresponding channel output can be modeled as

$$\begin{aligned} y(t) &= \int_{-\infty}^{\infty} h(t;\tau)x(t-\tau)d\tau \\ &= \int_{-\infty}^{\infty} H(t;f)X(f)e^{j2\pi ft}df, \end{aligned} \tag{3.1}$$

where $H(t;f)$ is the Fourier transform of the channel response at time t. Further consider that the spectral bandwidth W of the signal $X(f)$ is much smaller than the coherence bandwidth of the channel. In this case, the fading profile is known as flat fading. The coherence bandwidth is a statistical measure of the range of frequencies over which the channel response $H(t;f)$ remains approximately flat as a function of f. In other words, the coherence bandwidth is the range of frequencies over which two frequency components have a strong potential for amplitude correlation [48]. Mathematically, $|H(t;f_1)| \approx |H(t;f_2)|$ whenever $|f_1 - f_2|$ is less than the coherence bandwidth of the channel. Under the flat fading assumption, all the frequency components of $X(f)$ undergo the same attenuation and phase shift in transmission through the channel. This implies that, within the frequency band W occupied by $X(f)$, the time-varying transfer function $H(t;f)$ of the channel is essentially constant in the frequency variable. For such environments, the channel expression of (3.1) simplifies to

$$\begin{aligned} y(t) &= \int_{-\infty}^{\infty} h(t;\tau)x(t-\tau)d\tau \\ &\approx \int_{-\infty}^{\infty} h(t)e^{j\theta(t)}\delta(\tau)x(t-\tau)d\tau \\ &= h(t)e^{j\theta(t)}x(t). \end{aligned} \tag{3.2}$$

By construction, $H(t;f) = h(t)e^{j\theta(t)}$ over the frequency range of interest, $h(t)$ denotes the envelope attenuation and $\theta(t)$ represents the phase equivalent channel response. Thus, our frequency-nonselective (time selective) fading channel has a time-varying multiplicative effect on the transmitted signal. It is this latter model that we employ as a starting point for our analysis.

B. Wireless Channel

In addition to mean path attenuation and additive noise, which both affect data transmission over wireless channels, the channel model may include the time-varying filter of Section A. Before we present our mathematical model, we define the block fading channel introduced in [9], which will be used throughout our analysis. A block fading model represents a channel in which the connection statistics are fixed over constant-sized blocks. Such models are used to characterize slowly varying fading channel, which generally arise when the coherence time of the channel is relatively large compared to the delay constraint of the channel. The coherence time is a measure of the minimum period required for the magnitude change of the channel to become uncorrelated in time. We use block fading channels to model both channels in the tandem network. Unless otherwise specified, the block durations corresponding to both links is assumed to be identical, say T_{block}. The discrete-time channel is illustrated in Figure 7.

The transmitted signal $x(n)$ is subject to mean path attenuation $g(d)$, where d represents the distance between the mobile and its destination, multipath fading $h(n)$ and additive noise corruption $w(n)$. The signal at the destination for a coherent receiver can be represented by

$$y(n) = g(d)h(n)x(n) + w(n). \tag{3.3}$$

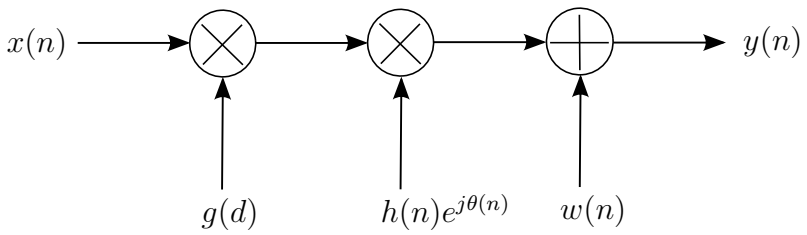

Fig. 7. Block diagram of a wireless channel.

The additive noise component $w(n)$ is often modeled as a proper complex white Gaussian process. Typically, the mean path attenuation $g(d)$ remains fixed over the timespan of interest. Therefore, the effect of $g(d)$ can be absorbed in the noise variance. Furthermore, $h(n)$ is normalized so that the expected power gain introduced by $h(n)$ is 1. In a heavily scattered environment, the fading process $h(n)$ is well-modeled as a zero-mean, proper complex Gaussian process. The envelope process $|h(n)|$ and the phase process $\angle h(n)$ form independent and stationary random processes, with $|h(n)|$ having a Rayleigh probability distribution function and the phase being uniform in $[0, 2\pi)$ [47]. Based on the normalization assumption, $|h(n)|$ has distribution

$$f(\xi) = 2\xi e^{-\xi^2}.$$

In general, a complete characterization of this random process requires that the higher-order statistics be specified in addition to the first-order statistics specified by the Rayleigh fading channel profile. For a wireless channel, the autocorrelation function of the fading envelope can be modeled using the zeroth order Bessel function of the first kind [49]. This function is derived under the assumption that the mobile terminal is moving in an isotropic environment at a constant velocity and is reasonable for short time intervals corresponding to movements of the order of few wavelengths. In the work described in this thesis, we model only the first-order statis-

tics of fading on encoded transmissions over both channels rather than specifying the higher-order statistics of $h(n)$. We model the wireless channels as discrete-time block fading process with i.i.d. Rayleigh distribution. Over block duration T_{block}, channel realizations remain the same; however, across blocks, they are independent and identically distributed. The reason for using this model is two-fold. First, we use block fading for the sake of mathematical tractability in solving the problem at hand. Second, we wish to take advantage of the independence structure introduced by the decoupling techniques described in Section B of Chapter II. We assume that the data is transmitted at a coderate R_i when $|h_i(n)|$ falls above a fixed threshold η_i, where the subscript $i = 1, 2$ represent the two hops. Otherwise, the data is lost and a new transmission is required for the packet to reach its destination. In reality, data is always transmitted at fixed rate R_i but success is achieved only when $|h_i(n)| > \eta_i$. Success/failure is determined by the ack/nack received at the transmitter from the receiver. Note that $n = 1, 2, \ldots$ represents the different blocks. The probability of $|h_i(n)|$ being above η_i can be obtained from the marginal Rayleigh distribution as

$$p_i = Pr\{|h_i(n)| \geq \eta_i\} = \int_{\eta_i}^{\infty} 2\xi e^{-\xi^2} d\xi = e^{-\eta_i^2}. \tag{3.4}$$

It follows that the transmission rate will be zero with probability $1 - p_i$.

C. Coding and Information Theory

Communication at the physical layer can be represented by parameters such as noise spectral density, bandwidth and transmit power. For additive white Gaussian noise (AWGN), the maximum rate at which error-free data transfers are possible is given by the Shannon capacity [1],

$$C = W \log_2 \left(1 + \frac{P}{N_0 W}\right) \quad \text{bits per second}, \tag{3.5}$$

where P is the power of the received signal, $N_0/2$ is the power spectral density of the noise process, and W is the spectral bandwidth. Recent developments in error-control coding allow operation close to channel capacity with minimal error-rates and small delays. The capacity expression of (3.5) can therefore be employed as an optimistic approximation of code performance. If a code is designed to operate at a rate R, the sent information can be decoded reliably if $R < C$, else it is lost.

A similar expression can be obtained for fading channels where the gain and hence the received power are time-varying in nature. Assuming the channel varies slowly over time, the instantaneous capacity of the wireless link is equal to

$$C_i(n) = W_i \log_2 \left(1 + \frac{|h_i(n)|^2 P_i}{N_0 W_i}\right) \quad \text{bits per second}, \quad i = 1, 2. \tag{3.6}$$

If the transmitted information is encoded at a rate R_i, it is assumed to reach its destination error-free provided that $R_i < C_i(n)$. On the other hand, if $R_i \geq C_i(n)$ then the information is lost as depicted in Figure 8 . This simplified characterization, which we use throughout for mathematical convenience, is valid provided that there are enough degrees of freedom available to allow the use of sophisticated codes during each data transfer. This model can be altered to accommodate practical codes and probabilities of link failures.

When channel state information is not known at the transmitter, a wireless node must transmit its data at a pre-selected coderate R_i to its destination. The threshold on the channel envelope can be derived based on the coderate used and on the condition that data is decoded only if $R_i < C_i(n)$ as in (3.6). This yields

$$|h_i(n)| > \eta_i = \sqrt{\frac{N_0 W_i}{P_i}\left(2^{\frac{R_i}{W_i}} - 1\right)} \quad i = 1, 2. \tag{3.7}$$

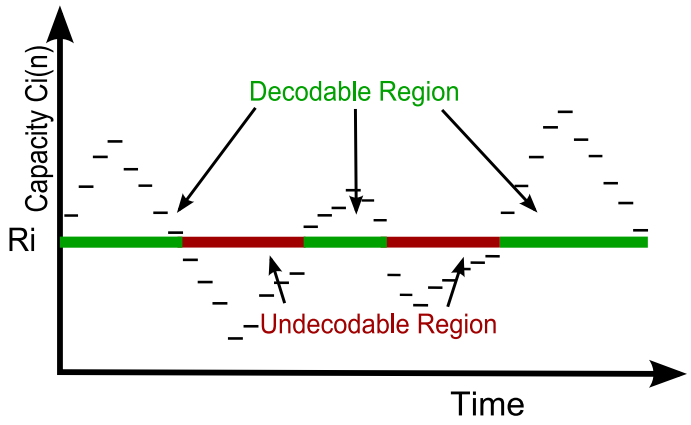

Fig. 8. Condition for information to be reliably decoded.

Once R_i is fixed, the average throughput of this channel is given by

$$R_i \Pr\{|h_i(n)| > \eta_i\} = R_i e^{-\eta_i^2}.$$

The maximum average throughput can be obtained by maximizing this expression over all admissible values of R_i. An immediate tradeoff can be observed between the likelihood of the channel being in good state and the rate at which data is transferred from the source to the destination. If we choose a large coderate R_i, the instantaneous throughput is high whenever the channel is in its good state. This indicates that we should choose a large threshold η_i. Yet, as η_i becomes larger, the probability that the channel remains in its good state decreases. Thus, the amount of time spent in the good channel state is less, and the average throughput suffers. On the other hand, if the coderate is decreased, the probability of being in the good state increases but the rate at which data is transmitted decreases.

We assume that a simple Automatic Repeat-reQuest (ARQ) scheme is in place at the physical layer, which informs the transmitter if the data has been decoded successfully or not. This behavior will affect the queue length distribution at the

terminal and, consequently, it will have a significant impact on performance. If the acknowledgments are received fast enough, the service offered to the user will be the same as the transmission rates supported by the wireless link. Hence, in our model, the corresponding service rate is R_i when the channel envelope exceeds the threshold and it is zero otherwise.

D. Queuing Model

It can be inferred from the decoupling techniques introduced in Section B of Chapter II that the upper bound on the overall performance of the two-hop system can be approximated by studying the two queues independently. The effective capacities of each queue will be calculated separately and conditions will be derived for the appropriate allocation of system resource. The underlying goal is to maximize the throughput of the system subject to the desired QoS constraint. Before we proceed with the performance analysis of the queues, we describe the actual discrete-time queue models used.

1. First Queue

Consider the first queue to be a simple discrete-time queue with a single server. We assume the data arrives at this buffer at a constant arrival rate a and gets served at a rate dependent on the realization of the wireless channel. Let $a_1(n)$ be a random variable denoting the number of bit arrivals in the nth block and $s_1(n)$ be another random variable denoting the number of bits served during that block. Thus, $a_1(n) = aT_{\text{block}}$ for all n; and $s_1(n)$ can be either zero or R_1T_{block} with probability $1 - p_1$ and p_1, as defined in (3.4). Let $q_1(n)$ be the length of the queue at time nT_{block}. Starting with an empty buffer being served under a work-conserving policy, we can write the

dynamic evolution of buffer as the Lindley's equation,

$$q_1(n+1) = (q_1(n) + aT_{\text{block}} - s_1(n))^+ \tag{3.8}$$

where $(x)^+ \triangleq \max(0, x)$. The amount of service offered in the interval $[n_1 T_{\text{block}}, n_2 T_{\text{block}})$ is given by

$$S_1(n_1, n_2) = \sum_{m=n_1}^{n_2-1} s_1(m). \tag{3.9}$$

For this queue to be stable, the expected arrival should be less than the expected service, i.e., $a < R_1 p_1$.

2. Second Queue

Let the second user also be represented by a simple discrete-time queue with one server. The arrival rate at this queue is the same as the service rate of the first channel, and the bits in this queue get served at a rate that depends on the realization of the second channel. In other words, the number of bit arrivals at discrete time n, $a_2(n)$ is considered to be the same as the service of the first queue $s_1(n)$ based on the decoupling principle. That is, $a_2(n)$ can be either zero or $R_1 T_{\text{block}}$ with probability $1 - p_1$ and p_1, respectively; it can be modeled as an on-off source. The amount of data served by the second queue at block n, $s_2(n)$ can be either zero or $R_2 T_{\text{block}}$ with probability $1 - p_2$ and p_2, as defined in (3.4). Let $q_2(n)$ be the length of the queue at time nT_{block}. Under assumptions similar to those made on the first queue, growth of the second queue is governed by the equation,

$$q_2(n+1) = (q_2(n) + a_2(n) - s_2(n))^+. \tag{3.10}$$

The cumulative arrival function over interval $[n_1 T_{\text{block}}, n_2 T_{\text{block}})$ is given by

$$A_2(n_1, n_2) = \sum_{m=n_1}^{n_2-1} a_2(m). \tag{3.11}$$

The amount of service offered in the interval $[n_1 T_{\text{block}}, n_2 T_{\text{block}})$ is equal to

$$S_2(n_1, n_2) = \sum_{m=n_1}^{n_2-1} s_2(m). \tag{3.12}$$

For this queue to be stable, we need $R_1 p_1 < R_2 p_2$.

CHAPTER IV

PERFORMANCE ANALYSES

To capture the impact of the physical layer parameters on the delay requirements of the system, we need to formulate an expression that will account for both aspects. Since we have already established that the effective capacity is the maximum input rate at which a system can operate subject to a given QoS constraint, we will use this concept to develop design criteria for the overall system. Using the channel and queueing models described in Section III, we will derive expressions for effective capacity and effective bandwidth and use them for further analysis to understand the allocation of resources as a function of the QoS parameter θ.

A. Queueing Performance Analysis

1. Effective Capacity of Outflow of First Queue

Based on the asymptotic probability of buffer overflow [5], the effective capacity of the first queue can be written as

$$\alpha_1(\theta) = \sup\left\{a : \lim_{x\to\infty} \frac{\log \Pr\{L_1 > x\}}{x} \leq -\theta\right\} \tag{4.1}$$

where L_1 is the steady state queue length. The effective capacity can also be written as

$$\alpha_1(\theta) = -\frac{\Lambda_1(-\theta)}{\theta} \tag{4.2}$$

provided $\Lambda_1(-\theta)$ satisfies the conditons of Gärtner-Ellis theorem [32], i.e., $\Lambda_1(-\theta)$ exists and is differentiable for all $\theta > 0$. $\Lambda_1(-\theta)$ is known as the asymptotic log-

moment generating function of the accumulated service and is defined as

$$\Lambda_1(-\theta) = \lim_{t\to\infty} \frac{1}{t} \log M(-\theta) \tag{4.3}$$

where $M(-\theta)$ is the moment generating function of $S_1(n_1, n_2)$ defined in (3.9),

$$M(-\theta) = E\left[e^{-\theta S_1(n_1,n_2)}\right]. \tag{4.4}$$

For a stationary ergodic sequence $\{s_1(n), n > 0\}$,

$$M(-\theta) = E[e^{-\theta S_1(1,n+1)}] \tag{4.5}$$

where $n = n_2 - n_1$. Substituting (4.5) in (4.3), we get

$$\Lambda_1(-\theta) = \lim_{n\to\infty} \frac{1}{nT_{\text{block}}} \log E\left[e^{-\theta S_1(1,n+1)}\right].$$

Since $s_1(n)$ is a sequence of i.i.d. random variables,

$$\Lambda_1(-\theta) = \frac{1}{T_{\text{block}}} \log E\left[e^{-\theta s_1(1)}\right]. \tag{4.6}$$

As observed, $\Lambda_1(-\theta)$ exists and is bounded for bounded $s_1(1)$. It also differentiable for $\theta > 0$, therefore, the effective capacity of the queue can be written using (4.2),

$$\alpha_1(\theta) = -\frac{1}{\theta T_{\text{block}}} \log E\left[e^{-\theta s_1(1)}\right]. \tag{4.7}$$

Based on (3.4) and the definition of $s_1(n)$ in Section D of Chapter III, it follows that

$$\alpha_1(\theta) = -\frac{1}{\theta T_{\text{block}}} \log\left(\left(1 - e^{-\eta_1^2}\right) + e^{-\eta_1^2} e^{-\theta R_1 T_{\text{block}}}\right). \tag{4.8}$$

The queue length $q_1(n)$ will be bounded exponentially with decay rate θ, if $a < \alpha_1(\theta)$. This will be explained in detail in Subsection 4.

2. Effective Bandwidth of Inflow of First Queue

Since $a_2(n)$ is a sequence of i.i.d. random variables, it is known from [6, 50] that the minimum envelope rate of $\{a_2(n), n > 0\}$ is the same as its effective bandwidth $\beta_2(\theta)$. Minimum envelope rate is the average rate of the minimum envelope process of the arrival process $\{a_2(n), n > 0\}$ as defined in [6]. Therefore, we can write

$$\beta_2(\theta) = \frac{1}{\theta} \log E\left[e^{\theta a_2(1)}\right]. \tag{4.9}$$

or, equivalently,

$$\beta_2(\theta) = \frac{1}{\theta T_{\text{block}}} \log\left(\left(1 - e^{-\eta_1^2}\right) + e^{-\eta_1^2} e^{\theta R_1 T_{\text{block}}}\right). \tag{4.10}$$

3. Effective Capacity of Outflow of Second Queue

The effective capacity of the service at the second queue can be obtained in a manner similar to the derivation of the effective capacity of the first queue,

$$\alpha_2(\theta) = -\frac{1}{\theta T_{\text{block}}} \log\left(\left(1 - e^{-\eta_2^2}\right) + e^{-\eta_2^2} e^{-\theta R_2 T_{\text{block}}}\right). \tag{4.11}$$

Even here the queue length $q_2(n)$ will be bounded exponentially with θ, if $\beta_2(\theta) < \alpha_2(\theta)$.

4. Probability of Buffer Overflow

As mentioned earlier, the probability of buffer overflow is an important performance metric that characterizes the behavior of a queue. Consider the first queue in the system. We will show that the tail of the steady state queue length distribution is bounded exponentially if the arrival rate is less than the effective capacity.

Expanding (3.8) recursively yields

$$\begin{aligned} q_1(n) = \max[0, aT_{\text{block}} - s_1(n-1), \ldots, \\ (n-1)aT_{\text{block}} - s_1(n-1) - s_1(n-2) - \ldots - s_1(1)]. \end{aligned} \tag{4.12}$$

Using the fact that $\max(x_1, x_2) \leq x_1 + x_2$ for $x_1, x_2 \geq 0$, we get

$$E\left[e^{\theta q_1(n)}\right] \leq \sum_{m=0}^{n-1} E\left[e^{\theta(maT_{\text{block}} - S_1(n-m,n))}\right] \tag{4.13}$$

The probability of buffer overflow is related to the moment generating function $E\left[e^{\theta q_1(n)}\right]$ through the Chernoff bound, as described in Section A of Chapter II, by

$$\Pr(q_1(n) \geq x) \leq e^{-\theta x} E\left[e^{\theta q_1(n)}\right]. \tag{4.14}$$

We can then write

$$\begin{aligned} \Pr(q_1(n) \geq x) &\leq e^{-\theta x} \sum_{m=0}^{n-1} E\left[e^{\theta(maT_{\text{block}} - S_1(n-m,n))}\right] \\ &\leq e^{-\theta x} \sum_{m=0}^{n-1} e^{\theta maT_{\text{block}}} E\left[e^{-\theta S_1(n-m,n)}\right]. \end{aligned}$$

Since $s_1(n)$ is a sequence of i.i.d. random variables,

$$\begin{aligned} \Pr(q_1(n) \geq x) &\leq e^{-\theta x} \sum_{m=0}^{n-1} e^{\theta maT_{\text{block}}} \left(E\left[e^{-\theta s_1(1)}\right]\right)^m \\ &\leq e^{-\theta x} \sum_{m=0}^{n-1} \left(E\left[e^{\theta(aT_{\text{block}} - s_1(1))}\right]\right)^m. \end{aligned}$$

As n approaches infinity, we have

$$\Pr(q_1(\infty) \geq x) \leq \frac{e^{-\theta x}}{1 - E\left[e^{\theta(aT_{\text{block}} - s_1(1))}\right]}, \tag{4.15}$$

provided $E\left[e^{\theta(aT_{\text{block}} - s_1(1))}\right] < 1$, which we know from (4.7) means $a < \alpha_1(\theta)$. A similar derivation for the second queue can show that the steady-state queue distribution

will be bounded exponentially with parameter θ provided that its effective bandwidth is less than its effective capacity.

B. Effective Capacity Analysis

The effective capacity quantifies the maximum supported arrival rate for a set of system parameters and a QoS constraint $\theta > 0$. It is an appropriate tool to quantify the optimal operating point of a delay-sensitive wireless system. This maximum rate can either be the true rate of a constant source or the effective bandwidth of a time-varying source. Our main objective is to find a, R_1 and R_2 for this optimal operating point. The effective capacity of a system with multiple queues in tandem is less than or equal to the effective capacity of the weakest link [51]. This fact can be confirmed from the large deviations principle as well. Observing expressions (4.8) and (4.11) and knowing that $R_1 < R_2$ insures stability, it is obvious that $\alpha_1(\theta) < \alpha_2(\theta)$. The objective is therefore to maximize the effective capacity of the first queue.

Table I. System Parameters.

$N_0 = 10^{-7}$ W/Hz	Noise power spectral density
$W_1 = W_2 = 10$ MHz	Bandwidth
$P_1 = P_2 = 100$ mW	Received power
$T_{\text{block}_1} = T_{\text{block}_2} = 2$ ms	Duration of block fade

Due to the complexity of the equations derived in Section A, it is hard to solve them manually. Hence, we will computationally solve for the arrival and service rates with the parameters of the wireless channel that appear in Table I as input. Mathematically, the objective function can be written as $\max \alpha_1(\theta)$ subject to the

condition $\beta_2(\theta) \leq \alpha_2(\theta)$ for a fixed θ.

The algorithm used to achieve this objective is as follows:

1. Choose $R_2 = \arg\max \alpha_2(\theta)$.
2. Find range of R_1 such that $\beta_2(\theta) \leq \alpha_2(\theta)$.
3. From this range, choose $R_1 = \arg\max \alpha_1(\theta)$.
4. Set $a = \max \alpha_1(\theta)$ because a is the maximum arrival rate that can be supported for a given θ.

Figure 9 shows the maximum supported arrival rate $\alpha_1(\theta)$ as a function of the QoS constraint θ for the system parameters in Table I. The figure also includes the optimal coderates R_1 and R_2 as a function of θ. The effective capacity at zero

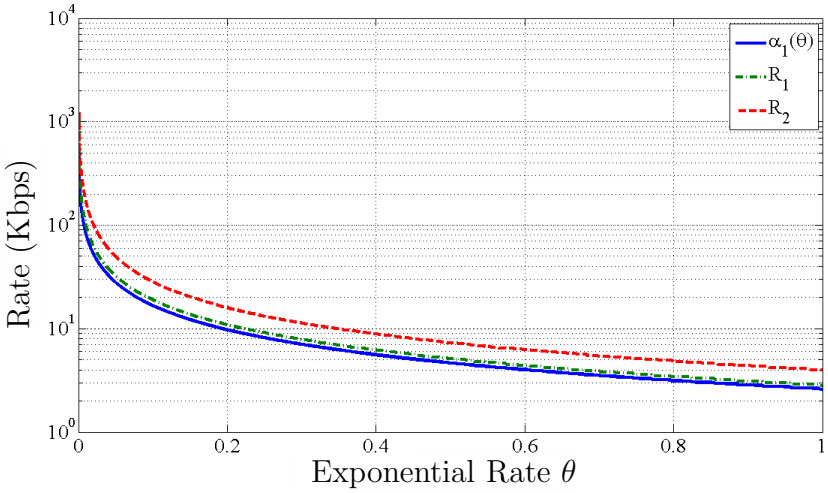

Fig. 9. Optimal coderates and effective capacity as a function of decay rate θ.

corresponds to the maximum throughput of the system. As the constraint value θ increases, the effective capacity decreases rapidly for fixed system parameters. This is intuitive since a lower arrival rate reduces the expected queue length. We also observe

that the optimal coderates R_1 and R_2 are functions of the QoS requirements. This is an important result which tells us that under strict QoS constraints, error control codes with lower rates perform better as they reduce the probability of the channel being OFF. This analysis offers a systematic way to select coderates as a function of channel profile and the QoS requirement of the system.

Another analysis can be performed to identify the dependence of the maximum arrival rate on the different bandwidth allocation among the channels. The bandwidth of the entire system is assumed to be 10 MHz. A fraction of the bandwidth γ is allocated to the first wireless channel, while the remaining $1-\gamma$ to the second channel. The other system parameters are selected according to Table I. The same algorithm as described earlier is applied to obtain different values of a for each γ by fixing θ. Figure 10 illustrates the results obtained from this analysis. For higher QoS

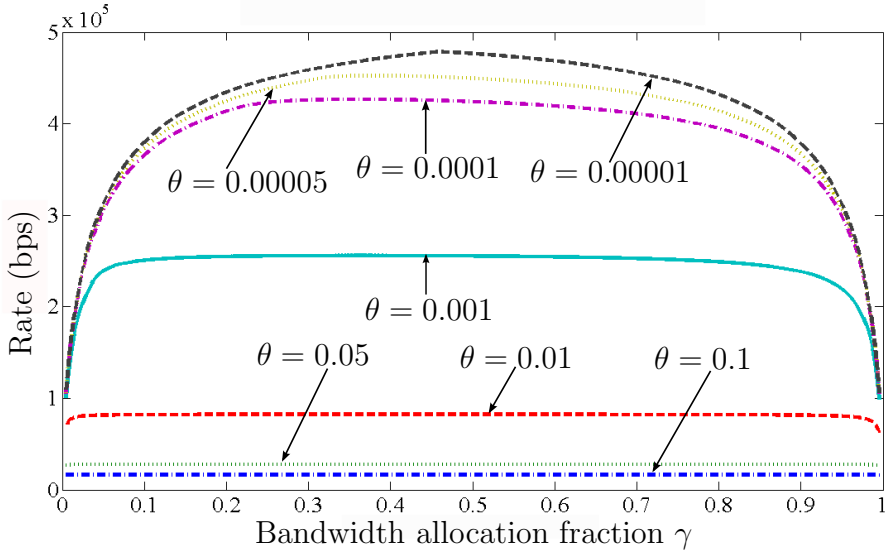

Fig. 10. Maximum arrival rate as a function of γ for various values of θ (symmetric links).

constraints, the maximum arrival rate remains almost constant for a wide range of γ

values. This suggests that the system hits the wide-band regime quickly under higher QoS constraints. Whereas for lower θ values, there is a significant difference in the maximum arrival rates as γ varies in $[0-1]$. When the system operates under such QoS constraints, the optimal operating point is observed to occur when γ lies in the interval $[0.4-0.5]$. This analysis provides a good point of operation for systems where delay limitation is less of an issue. Nevertheless, it reiterates the fact that allocation of fraction of bandwidth is not a predominant issue in delay-sensitive networks as long as the bandwidth for the entire system falls within the wide-band regime.

We also consider the cases in which there is a power imbalance between the two links. For the first case, we assumed $P1 = 100$ mW and $P2 = 25$ mW; for the second case, $P1 = 25$ mW and $P2 = 100$ mW. Analytical performance results were obtained by varying the bandwidth allocation fraction for different values of θ. Figures 11 & 12 display the results for both cases. As observed in the symmetric channel case, the

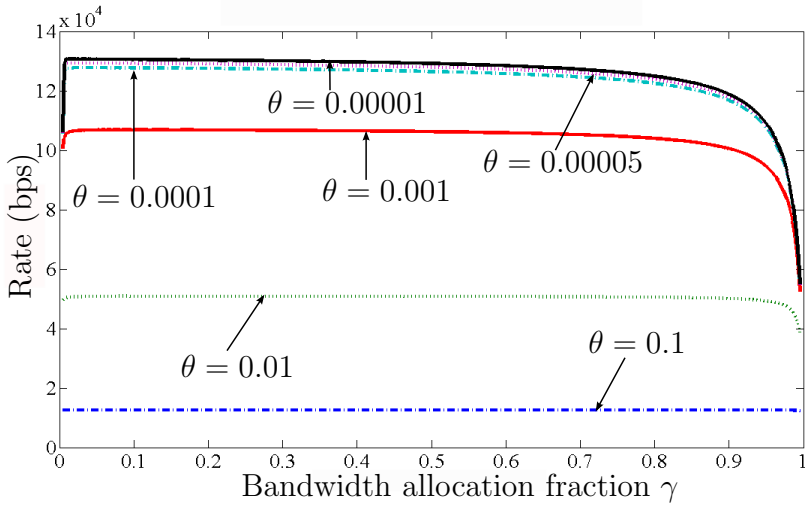

Fig. 11. Maximum arrival rate as a function of γ for different values of θ ($P1 > P2$).

effective capacity remains invariant to changes in the bandwidth allocation fraction for tighter QoS constraints. For lower values of θ, significant drop in the effective

capacity of the system is observed when a majority of the bandwidth is allocated to the stronger link. Apart from that, it is still almost constant for different bandwidth allocation policies.

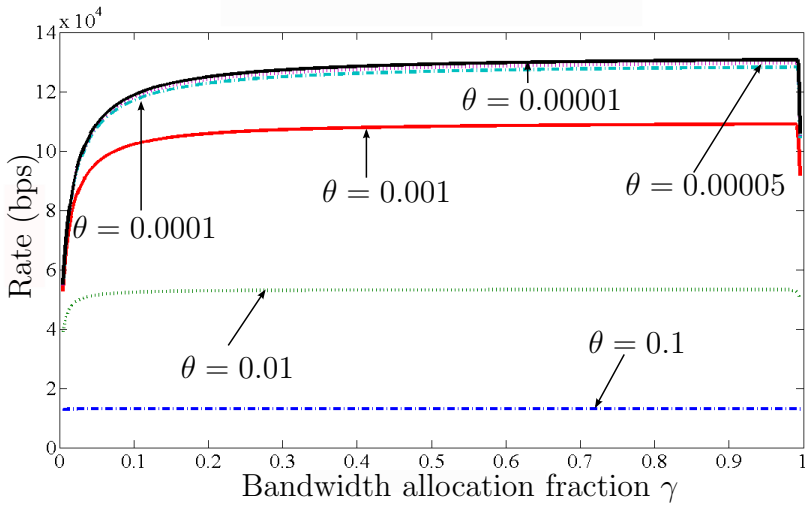

Fig. 12. Maximum arrival rate as a function of γ for different values of θ ($P1 < P2$).

Comparing the results from the three different cases, it can be inferred that spectral allocation doesn't really affect the performance of the system, provided the entire bandwidth is not allocated to either of the two channels. It should be noted that the effective capacity for the symmetric links with a particular service requirement is quite high compared to the links with power imbalances for the same service requirement. This is an expected result as the throughput of the entire system depends on the effective capacity of the weakest channel, which acts as a bottleneck.

CHAPTER V

SIMULATION RESULTS AND CONCLUSIONS

In addition to the numerical analysis carried in Chapter IV, we performed a simulation study for the two-hop system to support the analysis framework developed earlier. Let us consider a VoIP application where the maximum delay that can be tolerated is 300 ms. We simulate the probability of buffer overflow of the system for multiple QoS constraint values by specifying different arrival and service rates obtained from the analytical study as shown in Figure 9. To make a fair comparison, we use the system parameters of Table I to simulate the buffer overflow probability by allocating fraction of the bandwidth for both the wireless channels. The results are depicted in Figure 13. As expected the probability of overflow is almost zero for higher QoS constraints. This

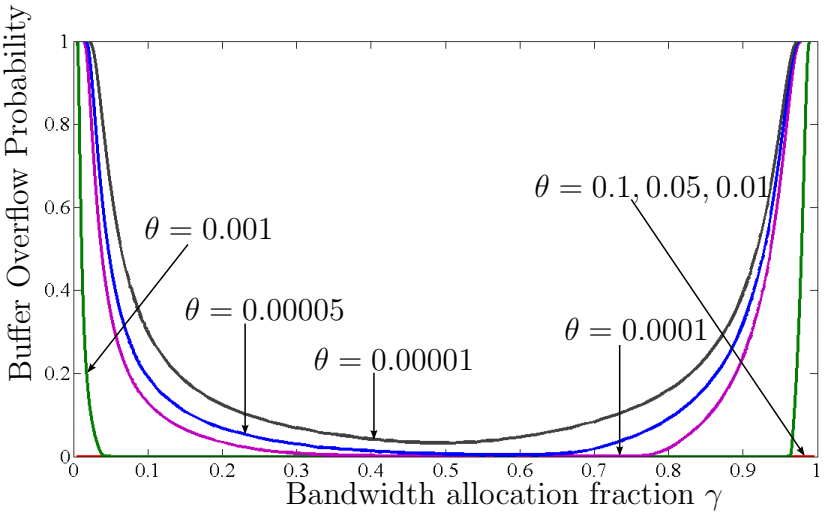

Fig. 13. Probability of buffer overflow as a function of bandwidth allocation fraction γ for different values of a, R_1 and R_2 (symmetric links).

is due to the fact that the buffer never gets filled up given the low rates supported by the system. For lower constraint values, the simulated system performs better within the γ-range $[0.4 - 0.5]$, which substantiates the results displayed in Figure 10.

Simulation scenarios that match the analysis scheme for wireless channels with imbalanced power constraints are also examined. The results of which are depicted in Figures 14 & 15. As seen earlier, the system is somewhat robust against spectral

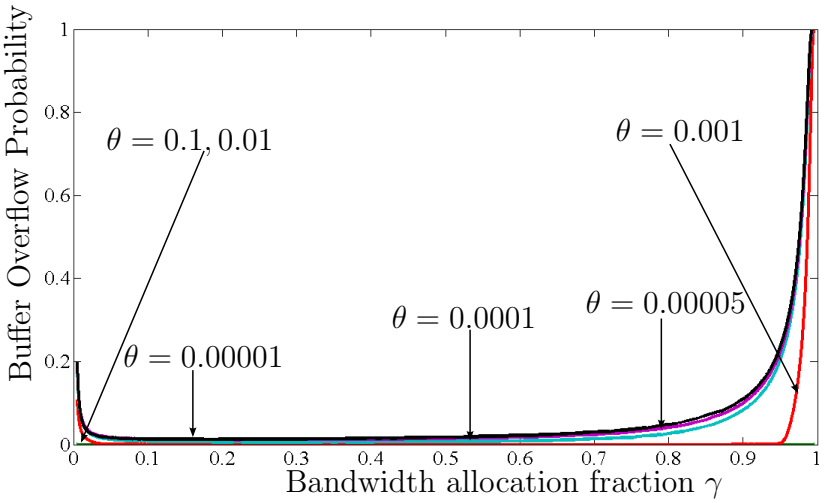

Fig. 14. Probability of buffer overflow as a function of bandwidth allocation fraction γ for different values of a, R_1 and R_2 ($P1 > P2$).

allocation for all values of θ. When θ has a high value, the rates supported by the system are so low that the buffers never exceed the threshold that satisfy the maximum delay allowed. The queues appear to be congested for the maximum time only when most of the bandwidth is allocated to the channel with higher power. These simulation results confirm the findings obtained via analytical methods and, thereby, justifying the cross-layer framework used for analyzing the two-hop system.

A. Conclusions

We investigated the performance of multihop wireless communication systems for delay-sensitive applications. We introduced decoupling techniques which provide us simpler ways to handle the tandem queues by studying them independently and yet

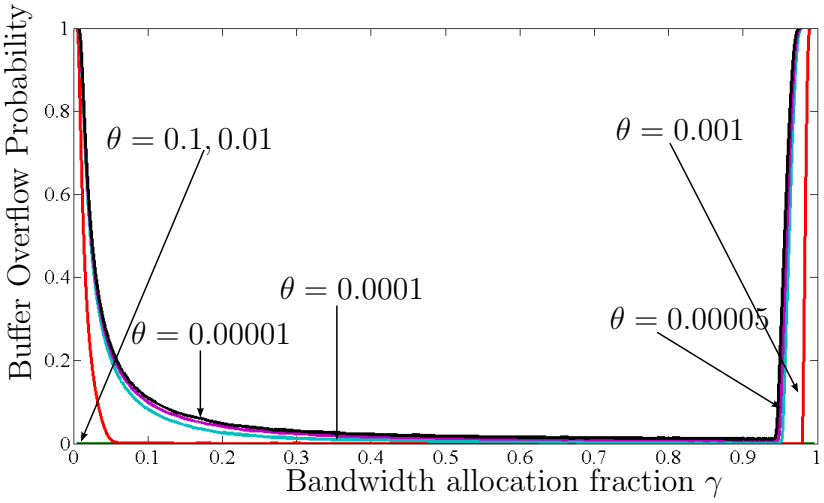

Fig. 15. Probability of buffer overflow as a function of bandwidth allocation fraction γ for different values of a, R_1 and R_2 ($P1 < P2$).

enable us to analyze the overall system performance. The wireless channel was modeled using i.i.d. Rayleigh block fading model to maintain the tractability in analyzing the system as well as to take advantage of the independence structure of the channel.

The system performance was evaluated using the large deviation principle governing the probability of buffer overflow, which is given by

$$-\lim_{x\to\infty}\frac{\log \Pr(L > x)}{x} = \theta.$$

This QoS metric is related to the concept of effective capacity which is defined as the maximum constant arrival rate that can be supported under a specific QoS constraint. The overall effective capacity for a system with multiple queues in tandem is dominated by the queue having the least effective capacity. When $\theta = 0$, the effective capacity approaches the maximum throughput. When the service constraints become more and more stringent, the effective capacity decays rapidly as a function of θ. Optimal code rates also depend heavily on the service requirement of the underlying application.

In this work, we studied the behavior of the effective capacity as a function of the fraction of total spectral bandwidth for a two-hop system with different power constraints for various service requirements. Overall performance of the tandem networks with balanced and imbalanced power requirements on each link were investigated. In the wideband regime, the effective capacity remains constant for wide range of bandwidth allocation fractions under strict QoS constraints. But for lower constraint values, there is significant difference in the effective capacity as the majority of the bandwidth is distributed to either of the two links. Overall, the system appears to be quite robust against spectral allocation schemes. Simulation results for the probability of buffer overflow also provide similar inferences and hence substantiate the fact that an analytic framework rooted in queueing theory may provide good insight for system design in the context of delay-sensitive communication systems.

Another important result to note is that the effective capacity decays rapidly as θ becomes large, which suggests that it is very difficult to support delay sensitive communication over wireless channel in the absence of channel state information. When channel knowledge is available, sophisticated power allocation schemes can be employed to dampen the rate of decay of the effective capacity. Our model did not incorporate these techniques to reduce its complexity in analyzing the performance of the system.

B. Scope of Future Work

This work can be extended to system with hop-counts greater than two. Different channel models can also be incorporated to study the effects of queueing behavior on the allocation of physical resources.

REFERENCES

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley-Interscience, 1991.

[2] S. Verdu and S. Shamai, "Spectral efficiency of cdma with random spreading," *IEEE Transactions on Information Theory*, vol. 45, no. 2, pp. 622–640, March 1999.

[3] S. Verdu, "Spectral efficiency in the wideband regime," *IEEE Transactions on Information Theory*, vol. 48, no. 6, pp. 1319–1343, June 2002.

[4] R. A. Berry and E. M. Yeh, "Cross-layer wireless resource allocation," *IEEE Signal Processing Magazine*, vol. 21, no. 5, pp. 59–68, September 2004.

[5] A. I. Elwalid and D. Mitra, "Effective bandwidth of general markovian traffic sources and admission control of high speed networks," *IEEE/ACM Transactions on Networking*, vol. 1, no. 3, pp. 329–343, August 1993.

[6] C.-S. Chang, "Stability, queue length, and delay of deterministic and stochastic queueing networks," *IEEE Transactions on Automatic Control*, vol. 39, no. 5, pp. 913–931, May 1994.

[7] C.-S. Chang and J. A. Thomas, "Effective bandwidth in high-speed digital networks," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 6, pp. 1091–1100, August 1995.

[8] R. A. Berry and R. G. Gallager, "Communication over fading channels with delay constraints," *IEEE Transactions on Information Theory*, vol. 48, no. 5, pp. 1135–1149, May 2002.

[9] L. H. Ozarow, S. Shamai, and A. D. Wyner, "Information theoretic considerations for cellular mobile radio," *IEEE Transactions on Vehicular Technology*, vol. 43, no. 2, pp. 359–378, May 1994.

[10] S. V. Hanly and D. Tse, "Multiaccess fading channels. ii: Delay-limited capacities," *IEEE Transactions on Information Theory*, vol. 44, no. 7, pp. 2816–2831, November 1998.

[11] W. Wu, A. Arapostathis, and S. Shakkottai, "Optimal power allocation for a time-varying wireless channel under heavy-traffic approximation," *IEEE Transactions on Automatic Control*, vol. 51, no. 4, pp. 580–594, April 2006.

[12] E. M. Yeh and R. A. Berry, "Throughput optimal control of cooperative relay networks," *IEEE Transactions on Information Theory*, vol. 53, no. 10, pp. 3827–3833, October 2007.

[13] C.-S. Chang, *Performance Guarantees in Communication Networks.* Telecommunication Networks and Computer Systems, New York: Springer, 1995.

[14] D. Wu and R. Negi, "Effective capacity: A wireless link model for support of quality of services," *IEEE Transactions on Wireless Communication*, vol. 2, no. 4, pp. 630–643, July 2003.

[15] D. Wu and R. Negi, "Downlink scheduling in a cellular network for quality-of-service assurance," *IEEE Transactions on Vehicular Technology*, vol. 53, no. 5, pp. 1547–1557, September 2004.

[16] D. Wu and R. Negi, "Utilizing multiuser diversity for efficient support of quality of service over a fading channel," *IEEE Transactions on Vehicular Technology*, vol. 54, no. 3, pp. 1198–1206, May 2005.

[17] L. Liu and J.-F. Chamberland, "On the effective capacities of multiple-antenna gaussian channels," Submitted to the *IEEE Transactions on Information Theory*, September 2007.

[18] L. Liu, P. Parag, J. Tang, W.-Y. Chen, and J.-F. Chamberland, "Resource allocation and quality of service evaluation for wireless communication systems using fluid models," *IEEE Transactions on Information Theory*, vol. 53, no. 5, pp. 1767–1777, May 2007.

[19] R. L. Cruz, "A calculus for network delay, part i: Network elements in isolation," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 114–131, January 1991.

[20] R. L. Cruz, "A calculus for network delay, part ii: Network analysis," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 132–141, January 1991.

[21] K. Romer and F. Mattern, "The design space of wireless sensor networks," *IEEE Wireless Communications*, vol. 11, no. 6, pp. 54–61, December 2004.

[22] Y. Hua, Y. Huang, and J. J. Garcia-Luna-Aceves, "Maximizing the throughput of large ad hoc wireless networks," *IEEE Signal Processing Magazine*, vol. 23, no. 5, pp. 84–94, September 2006.

[23] H. V. Poor, *An Introduction to Signal Detection and Estimation*. New York: Springer, 2nd edition, 1998.

[24] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume 2: Detection Theory*. New Jersey: Prentice Hall PTR, 1998.

[25] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume 1: Estimation Theory*. New Jersey: Prentice Hall PTR, 1993.

[26] A. Ephremides, "Energy concerns in wireless networks," *IEEE Wireless Communications*, vol. 9, no. 4, pp. 48–59, August 2002.

[27] R. Min, M. Bhardwaj, S.H. Cho, N. Ickes, E. Shih, A. Sinha, A. Wang, and A. Chandrakasan, "Energy-centric enabling technologies for wireless sensor networks," *IEEE Wireless Communications*, vol. 9, no. 4, pp. 28–39, August 2002.

[28] V. S. Raghunathan and C.S.P. Srivastava, "Energy-aware wireless microsensor networks," *IEEE Signal Processing Magazine*, vol. 19, no. 2, pp. 40–50, March 2002.

[29] R. R. Tenny and N. R. Sandell, "Detection with distributed sensors," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 17, no. 4, pp. 501–510, July 1981.

[30] R. Viswanathan and P. K. Varshney, "Distributed detection with multiple sensors part i - fundamentals," *Proceedings of the IEEE*, vol. 85, no. 1, pp. 54–63, January 1997.

[31] R. S. Blum, S. A. Kassam, and H. V. Poor, "Distributed detection with multiple sensors part ii - advanced topics," *Proceedings of the IEEE*, vol. 85, no. 1, pp. 64–79, January 1997.

[32] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications.* New York: Springer, 2nd edition, 1998, Stochastic Modeling and Applied Probability.

[33] P. G. Dupuis and R. S. Ellis, *A Weak Convergence Approach to the Theory of Large Deviations.* New York: Wiley-Interscience, 1997, Wiley Series in Probability and Statistics.

[34] J. N. Tsitsiklis, "Decentralized detection by a large number of sensors," *Mathematics of Control, Signals, and Systems (MCSS)*, vol. 1, no. 2, pp. 167–182, 1988.

[35] V. Anantharam, "A large deviations approach to error exponents in source coding and hypothesis testing," *IEEE Transactions on Information Theory*, vol. 36, no. 4, pp. 938–943, July 1990.

[36] R. S. Blum and S. A. Kassam, "On the asymptotic relative efficiency of distributed detection schemes," *IEEE Transactions on Information Theory*, vol. 41, no. 2, pp. 523–527, March 1995.

[37] P. Gupta and P. R. Kumar, "The capacity of wireless networks," *IEEE Transactions on Information Theory*, vol. 46, no. 2, pp. 388–404, March 2000.

[38] P. Gupta and P. R. Kumar, "Towards an information theory of large networks: an achievable rate region," *IEEE Transactions on Information Theory*, vol. 49, no. 8, pp. 1877–1894, August 2003.

[39] R. G. Gallager, *Information Theory and Reliable Communication.* New York, NY, USA: John Wiley & Sons, Inc., 1968.

[40] R. S. Ellis, *Entropy, Large Deviations and Statistical Mechanics.* New York: Springer-Verlag, 1985.

[41] A. Shwartz and A. Weiss, *Large Deviations for Performance Analysis.* London: Chapman and Hall, 1995.

[42] L. Liu, P. Parag, and J.-F. Chamberland, "Quality of service analysis for wireless user-cooperation networks," *IEEE Transactions on Information Theory*, vol. 53, no. 10, pp. 3833–3842, October 2007.

[43] P. Parag and J.-F. Chamberland, "Queuing analysis of butterfly network," Submitted to the *IEEE International Symposium on Information Theory*, January 2008.

[44] W. C. Jakes, *Microwave Mobile Communications*. New York: John-Wiley and Sons, 1974.

[45] R. S. Kennedy, *Fading Dispersive Communication Channels*. New York: Wiley Interscience, 1969.

[46] P. Viswanath and D. Tse, *Fundamentals of Wireless Communication*. New York: Cambridge University Press, 2005.

[47] E. Biglieri, J. Proakis, and S. Shamai, "Fading channels: Information theoretic and communication aspects," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2619–2692, October 1998.

[48] T. S. Rappaport, *Wireless Communications: Principles and Practice*. New Jersey: Prentice-Hall, 2nd edition, 2001.

[49] V. V. Veeravalli and A. Sayeed, "Wideband wireless channels: Statistical modelling, analysis and simulation," University of Illinois, Urbana-Champaign, 2004.

[50] F. P. Kelly, "Effective bandwidths at multi-class queues," *Queueing systems*, vol. 9, pp. 5–16, 1991.

[51] D. Wu and R. Negi, "Effective capacity-based quality of service measures for wireless networks," *Mobile Networks and Applications*, vol. 11, no. 1, pp. 91–99, February 2006.

VITA

Name: Omar Ahmed Ali

Address: Department of Electrical and Computer Engineering,
Texas A&M University,
214 Zachry Engineering Center,
College Station, Texas 77843-3128

Email Address: omar.a.ali@gmail.com

Education: B.Tech, Electrical Engineering,
Indian Institute of Technology Madras, 2005
M.S., Electrical Engineering, Texas A&M University, 2008